A Comprehensive Study of Network Compression for Image Classification

Jian Cheng

jcheng@nlpr.ia.ac.cn

NLPR & AIRIA
Institute of Automation, Chinese Academy of Sciences, China

Introduction

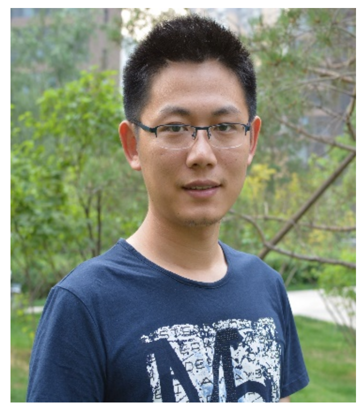
Team

- **Team Name**: RIAIR
- **Affiliation:** NLPR & AIRIA, Institute of Automation, Chinese Academy of Sciences, China
- **Team Member**:

Peisong Wang

Xiangyu He

Tianli Zhao



Cong Leng

Yifan Zhang

Jian Cheng

Introduction

MicroNet Challenge

- **MicroNet**
  - Aim: to build efficient models for the specified tasks with required conditions.
  - Tasks: ImageNet, CIFAR-100, WikiText-103
  - Criteria: Math operations, parameter Storage

Introduction

MicroNet Challenge

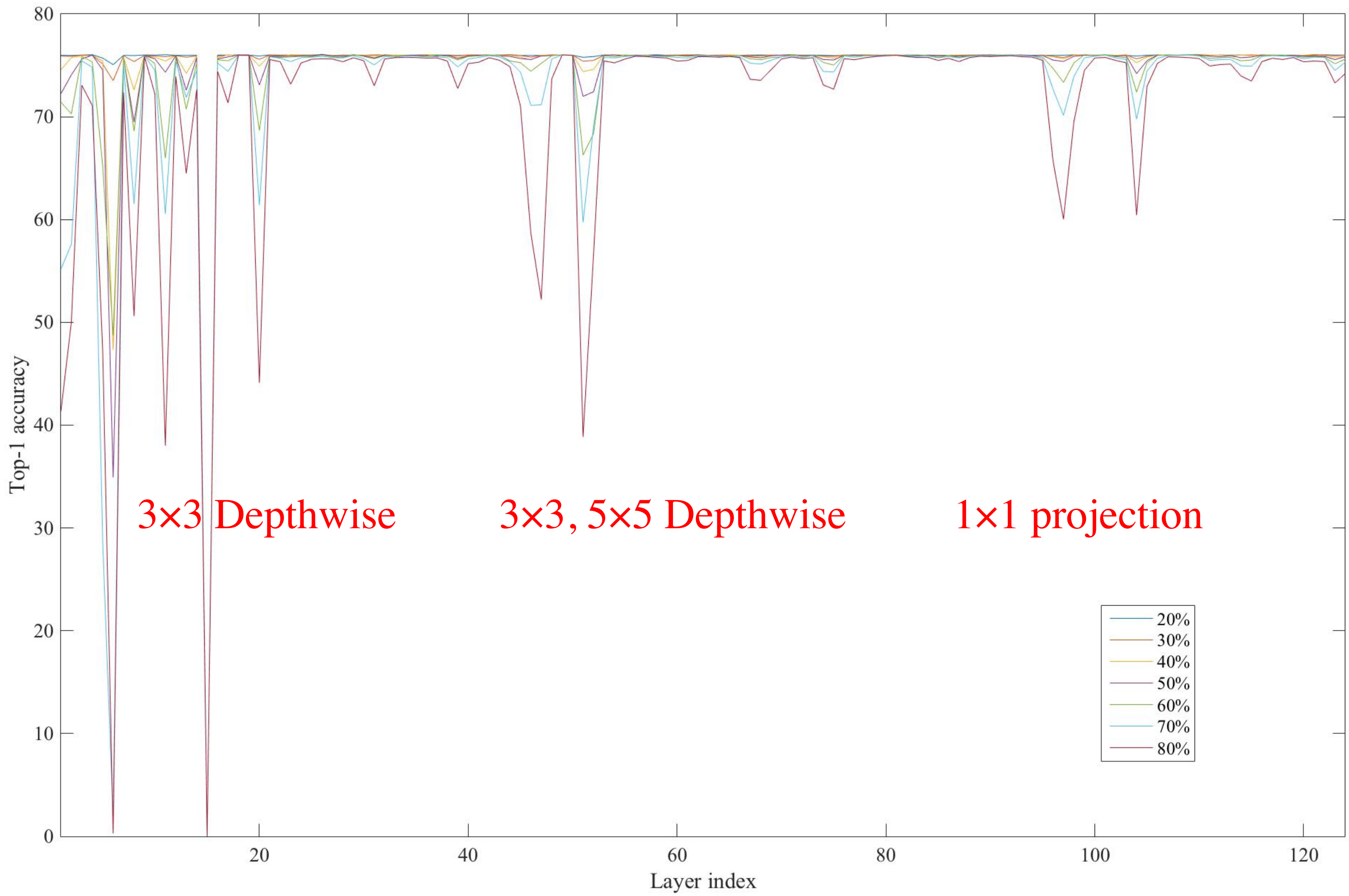
- **Task 1**: ImageNet Classification
  - 75% top-1 accuracy
  - Normalized relative to MobileNetV2-1.4
    - 6.9M parameters,
    - 1170M operations
  - Score = Storage / 6.9M + Ops/ 1170M
- **Task 2**: CIFAR-100 Classification
  - 80% top-1 accuracy
  - Normalized relative to WideResNet-28-10
    - which has 36.5M parameters
    - 10.49B math operations
  - Score = Storage / 36.5M + Ops/ 10.49B

Model Selection

- Select models with a bit higher accuracy than target quality
    - Automl searched models are more preferable
- MixNet for ImageNet Track (Searched on ImageNet)
    - MixNet-S for student (75.9%)
    - MixNet-L for teacher (78.9%)
- DenseNet for CIFAR-100 Track
    - DenseNet-100 for student (81.1%)
    - DenseNet-172 for teacher (84.0%)

Robust Analysis

- Different layers have different robustness to sparsity



Pruning

- Static Pruning
  - Set the smallest proportion of weights to zeros.
  - No grad to masked weights
  - **Fixed mask** during finetuning
- Dynamic Pruning
  - Pruned weights also get gradients
  - **Update mask** before the next SGD iteration
- Progressive Pruning
  - No grad to masked weights
  - **Update mask** before the next epoch

Pruning Results

- Static Pruning ~ Progressive Pruning > Dynamic Pruning

Static pruning results

Models	Param Sparsity	Op Sparsity	Top-1 Accuracy
Original	0	0	75.98
Sparse	58.6 %	45.9 %	75.57
Sparse, large layer ≥ 50 %	**59.6 %**	**47.1 %**	**75.56**
Sparse, large layer ≥ 60 %	63.4 %	59.3 %	75.09

Knowledge Distillation

- KD always improves accuracy
- Stronger teacher means higher accuracy?

Student	Teacher	Teacher Acc.	Top-1 Accuracy
MixNet-s-pruned	-	-	74.4 %
MixNet-s-pruned	MixNet-s	75.9 %	74.6 %
MixNet-s-pruned	MixNet-m	77.2 %	74.9 %
MixNet-s-pruned	**MixNet-l**	**78.9 %**	**75.0 %**
MixNet-s-pruned	SENet154	81.3 %	74.7 %

Quantization

- The same quantization for weights and activations
- Quantization Function:

$$q(x) = clamp(round(x/\alpha), Q_{min}, Q_{max})$$

$\alpha \in R$ is a scaling factor

- Activation quantization
  - Each layer has a scaling factor
- Weight quantization
  - Each kernel has a scaling factor

Quantization

- Problem:

$$Q \approx \alpha X$$

- Optimization:

$$min \parallel Q - \alpha X \|_F^2$$

- Iterative optimization
  - Given alpha, optimize Q

$$Q = q(x) = clamp(count(x/alpha), Q_{min}, Q_{max})$$

  - Given Q, optimize alpha

$$\alpha = \frac{X^T Q}{Q^T Q}$$

Activation Quantization

- Extract features using a batch of images
- Optimize scaling factors
- Finetune with fixed scaling factors

Models	Sparsity	Activation	Weight	Top-1 Accuracy
Sparse	59.6 %	FP32	FP32	75.56
Sparse_AQ8	59.6 %	8-bit	FP32	75.82
Sparse_AQ7	**59.6 %**	**7-bit**	**FP32**	**75.65**
Sparse_AQ6	59.6 %	6-bit	FP32	75.51

Weight Quantization

- Optimize scaling factors
- Finetune with fixed scaling factors

```
bitwidth = [7, 7, 7, 7, 7, 7, 7, 7, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 5, 5, 5, 5, 5,
            5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 5, 5, 5, 5, 5,
            7, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,
            5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 5, 5, 5, 5,
            5, 4, 4, 4, 4, 5, 5, 5, 5, 4, 4, 4, 4, 4, 5, 5, 5, 5, 3, 3, 4, 4, 3, 3]
```

Models	Sparsity	Activation	Weight	Top-1 Accuracy
Sparse	59.6 %	FP32	FP32	75.56
Sparse_AQ7	59.6 %	7-bit	FP32	75.65
Sparse_AQ7_WQ	**59.6 %**	**7-bit**	**7-5-4-3-bit**	**75.05**

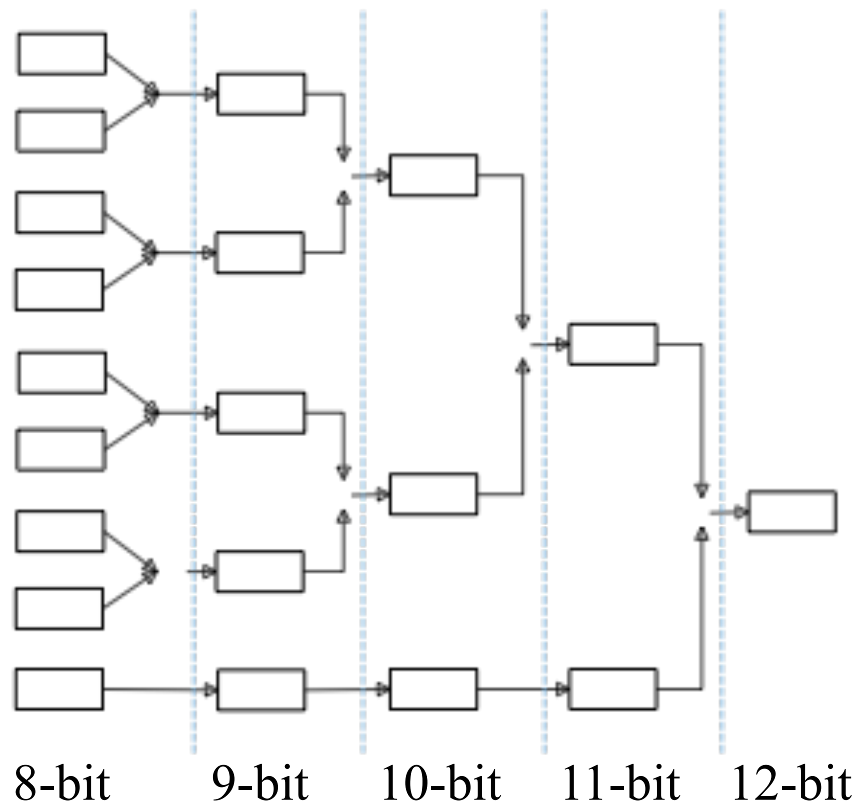
Scoring

- Low-bit quantization allows low-bit accumulation

1. Tree adder
   Different bitwidth for different levels of addition
   Max bitwidth for N elements:

$$L = l + \lceil log_2(N) \rceil$$

2. Integer adder
   Max bitwidth for all levels
3. FP16 adder
4. FP32 adder



Scoring

Final Results for ImageNet Classification

adder-type	op score	param score	final score
Tree	0.080077	0.049394	0.129463
Int	0.092668	0.049394	0.142063
FP16	0.088973	0.049394	0.138368
FP32	0.139955	0.049394	0.189349

0.34M parameters, 93.7M operations
20.2× compression, 12.5× acceleration

CIFAR-100 Classification Task

Model	Top-1 Accuracy
DensNet-172	84.00 %
DensNet-100	81.17 %
CONV-Prune-75%	81.01 %
Activation-4bit	80.24 %
CONV-Weight-4bit	80.28 %
FC-Prune-50%	80.38 %
FC-Weight-4bit	80.34 %

Final Results for CIFAR-100 Task

adder-type	op score	param score	final score
Tree	0.002805	0.001365	0.004169

49.8K parameters, 29.4M operations
732.6× compression, 356.5× accelerlation

Summary

- Proposed a network compression framework
  - Pruning
  - Quantization
- Practical tips
  - Robust analysis
  - Knowledge distillation
- The simplest method works good
  - With limited time!
- Extreme quantization
  - Binary/tenary
- Neural architecture search for composite compression

AiRiA
中国科学院自动化研究所
南京人工智能芯片创新研究院

CASIA

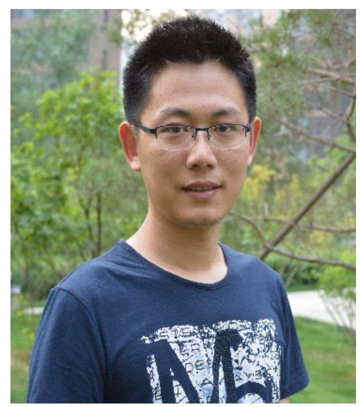
Thanks for all your attention!

Code: https://github.com/wps712/MicroNetChallenge

Peisong Wang

Xiangyu He

Tianli Zhao



Cong Leng

Yifan Zhang

Jian Cheng